计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211100063

一种专利知识图谱的构建方法

邓 亮[1,2,3] 曹存根[4]
1 中国科学院大学计算机科学与技术学院 北京 100049
2 中国科学院沈阳计算技术研究所 沈阳 110168
3 国家知识产权专利局 北京 100083
4 中国科学院计算技术研究所 北京 100190
(dengliang@cnipa.gov.cn)

摘 要 专利知识图谱对专利精准检索、专利深度分析和专利知识培训等应用起到了重要作用。文中提出了一种实用的基于种子知识图谱、文本挖掘以及关系补全的专利知识图谱构建方法。在该方法中,为确保质量,首先人工建立一个种子专利知识图谱,然后采用专利文本模式的概念和关系抽取方法扩展种子专利知识图谱,最后对扩展的专利知识图谱进行定量评估。文中针对中医药领域专利进行了种子知识的人工提取和词法句法模式的人工总结,并使用机器学习的方法在学习到新的词法句法模式后对种子专利知识图谱进行扩展和图谱补全。实验结果表明,中医药领域专利种子知识图谱中的节点数和关系数分别为19453个和194775条,经过扩展后,它们分别达到了558461个和7275958条,即分别增加了27.7倍和36.3倍。

关键词:专利文本;专利知识图谱;词法句法分析;表示学习
中图法分类号 TP391

Methods of Patent Knowledge Graph Construction

DENG Liang[1,2,3] and CAO Cun-gen[4]
1 School of Computer Science and Technology,University of Chinese Academy of Sciences,Beijing 100049,China
2 Shenyang Institute of Computing Technology,Chinese Academy of Sciences,Shenyang 110168,China
3 Patent Office,China National Intellectual Property Administration,Beijing 100083,China
4 Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190,China

Abstract Patent knowledge graph plays a important role in patent accurate retrieval,patent in-depth analysis and patent knowledge training. This paper proposes a practical patent knowledge graph construction method based on seed knowledge graph,text mining and relationship completion. In this method,to ensure the quality,a seed patent knowledge graph is first established manually,then the concept and relation extraction method of patent text pattern is used to expand the seed patent knowledge graph,and finally the extended patent knowledge graph is quantitatively evaluated. In this paper,artificial extraction of seed knowledge and manual summarization of lexical and syntactic patterns are carried out for patents in the field of traditional Chinese medicine. After obtaining new lexical and syntactic patterns by machine learning,the knowledge graph of seed patent is expanded and completed. Experimental results show that the number of nodes and relationships in the knowledge graph of traditional Chinese medicine are 19453 and 194775 respectively. After expansion,they reach 558461 and 7 275958 respectively,representing an increase of 27.7 and 36.3 folds respectively.

Keywords Patent text,Patent knowledge graph,Lexical and syntactic analysis,Representation learning

1 引言

专利指受到专利法保护的发明创造,即专利技术,是受国家认可并在公开的基础上进行法律保护的专有技术。中国的专利申请数量以每年超过10%的增速在持续增加。截至2022年,中国已有超过3750万数量的专利书[1]。

随着我国政府对自主创新能力的日益重视以及对专利战略的研究,我国的知识产权体系得到了快速发展,专利申请数量呈现指数型增长。根据2021年世界知识产权组织(WIPO)发布的统计数据,全球共有620万项专利申请,其中405万项来源于中国。全世界专利申请数量同比增长5.2%,其中中国专利申请增幅尤其大,达到了11.6%[1]。

专利作为人类发明成果的载体,是人类各领域科技人员智慧的结晶,包含着巨大的科学价值和行业知识,是企业和普通大众创新时获取相关信息与知识的重要途径之一。根据世界知识产权组织的统计,世界上90%95%的研发成果包含在专利文献中,其中约有80%的发明成果只通过专利文献公开,利用专利信息可缩短60%的研发时间,节

到稿日期:2021-11-05 返修日期:2022-03-11
通信作者:曹存根(cgcao@ict.ac.cn)

省 40%以上的研发费用。

专利文献数据库是高价值密度数据，如何从这样一个庞大的知识宝库中挖掘出深层次价值，一直都是知识产权领域企业和政府机构的科研目标。人们通过对专利数据的挖掘，实现了包括专利价值评估、专利相似度分析、企业专利布局、市场热点预测以及专利生产导航等不同方面的行业应用。随着这些应用种类的不断增多以及应用深度的加深，我们发现专利应用分析的难度越来越大，其主要原因之一是缺乏一个作为专利数据挖掘的底层基础知识库系统。这个知识库系统需要结合专利业务领域知识，对海量专利文献数据进行深层次的解析，从中提取出结构性的知识点，然后汇总、关联成为有一定体量的数据库系统，方便人们查询使用；同时，专利知识库系统还需要随着专利文献数据的不断增加而迭代扩展，以保证及时性和完整性。

知识图谱技术作为目前用于实现海量知识库系统的主流技术路线，满足了我们对专利知识库系统的所有需求。首先，在知识图谱构建过程中进行各种概念和实体的知识点提取，对这些知识点的提取就是对专利文献进行解析的过程；然后，需要对知识点之间的关系进行抽取，即对专利知识点进行梳理的过程；最后，将构建好的知识图谱以可视化形式展现出来，以便检索和查看。知识图谱具备成熟的新知识扩展机制，保障了知识库系统的迭代更新。

构建一个完整的专利文献知识图谱，简称专利图谱，有很高的应用价值，有助于实现众多的专利领域业务工作和数据挖掘目标。首先，专利图谱可以对专利审查工作起到提质、增效的作用。我国目前有大约 1.6 万名专利审查员，面对每年 400 多万篇专利的申请量，每个审查员平均每个工作日要完成至少一个案子的检索，而每个案子的检索都要与包括中文和外文专利文献在内的超过 1.1 亿篇文献进行比对，检索负担繁重，稍有疏忽，就会造成严重的漏检后果。而专利图谱作为一个知识库系统，已经在语义层面上提取并且包含了所有文献的知识点，对于认定新申请文献的新颖性和创造性有很大的辅助作用。相比传统的关键词和分类号检索方式，这种基于专利语义层面的检索具有更高的准确率；同时，专利图谱系统的构建都会有性能上的考虑，充分保障了检索效率。因此，构建专利文献知识图谱，有利于提升专利检索的质量和效率。

其次，一个高质量的专利图谱又是一个多领域知识库系统，有助于各行各业人才学习领域知识，并且产生新的发明创造。专利图谱具有知识点丰富以及可视化展示的特点，是广大学生、科研人员和企业研发人员学习前沿专业领域知识的有力工具之一。另外，专利图谱是对专利文献内容的提炼和梳理，有助于提取发明要素，总结不同领域的专利发明模式，给广大企业研发人员和科研工作者提供发明创新的思路和参考样例，起到促进发明创新的作用。

最后，上文提到的专利价值评估、专利相似度分析以及市场技术热点分析等应用场景，都需要以专利图谱中的知识点作为数据基础的重要组成部分来展开分析，从而实现更精准的分析目标。

综上所述，构建专利图谱的工作意义重大，本文的研究工作将围绕专利图谱构建工作进行。我们将基于中医药领域的大量专利文献构建中医药领域专利图谱。构建图谱的基本方法是：先根据专家知识和领域词典构建规模较小的种子知识图谱；然后通过人工和自动学习方式，发现知识点提取句法模式；最后基于种子知识图谱和知识点提取句法模式，并对其进行扩展，形成完备的中医药领域专利图谱。

2 相关工作

2012 年，Google 最早提出了知识图谱，他们随即开始利用知识图谱技术改善搜索引擎核心。知识图谱本质上是一种语义网络，它用图的形式描述客观事物，由结点和边组成。每个知识点表示一个三元组(Subject-Predicate-Object，SPO)，也可以记做 HRT(Head-Relation-Tail)。

已有研究使用专利信息在知识图谱构建方面进行了探索。2017－2018 年，Xu[2]，Xu[3]，Xun[4] 分别针对智能家居、抗肝癌药物以及云计算 3 个不同领域的专利文献进行了专利图谱的构建工作。这 3 个专利图谱所涉及的相关领域专利文献数量分别为 1 068，4 746，16 123，数据量相对较少。其中，Xu[2] 的专利图谱是智能家居领域的专利相似度图谱，图谱中使用专利对象作为实体结点，发现高相似度文献之间存在相似关系，实现了专利文献聚类的效果。Xu[3] 使用少量专利数据构建了抗肝癌药物领域关键词共现图谱和发明人关系图谱。Sun[4] 通过关键词提取构建了云计算领域专利的关键词共现图谱。上述工作均缺乏对专利内容在语义层面上的深度解析。

Zhang 等[5] 和 Gao[6] 分别在 2019 年和 2020 年进行了专利图谱构建工作。工作中所涉专利数量分别为 305 和 406，均是小规模图谱。前者构建了太子参相关医药领域的发明人关系图谱和关键词图谱，后者则构建了蒙医药领域的发明人关系图谱和关键词图谱。这两项工作涉及的专利数量较少，主要是受限于太子参和蒙医药领域的专利文献总体数量，对于大量专利文献的检索推荐工作尚无显著借鉴意义。

Zhang[7] 于 2019 年构建了全领域专利知识图谱，其涉及专利数量达 200 万以上，是目前包含中文专利最多的知识图谱，对专利检索有一定的促进作用。该专利图谱包含了专利之间的相似度关系、专利和发明人之间的申请关系以及专利和技术领域之间的 TPC 分类关系。该项工作中，作者采用 RAKE 和 K-means 方法相结合来实现专利实体识别后将实体存入数据库，并通过 Select 语句选取相应实体进行关系构建。文中没有提及该方法获取的知识图谱实体节点的有效性评价等相关工作。

国外方面，Sarica 等[8] 通过挖掘 1976 年以来申请的专利，构建了涵盖技术领域基本概念的知识图谱，并将其作为基础架构以支持其他专利应用。其通过学习大量专利文献，提取了不同领域技术词汇之间的相关度，但没有根据专利文本的特殊结构针对专利内容本身进行深层次的解析。

综合以上比较分析，目前业内的专利图谱大部分聚焦在专利文献和发明人之间的所属关系上，还有部分图谱在单词统计层面上做了领域关键词相关度分析工作。但它们都没有实现对专利内容在语义层面上的深度解析和知识点的深度挖掘；同时，大部分现有专利图谱由于受到研究目的的局限，规模较小，仅使用人工方法构建即可满足应用需要，在大规模

专利应用中缺乏实际使用价值。Sun[4]和 Zhang[7]提出的知识图谱虽涉及相对较多的专利数据，但没有涉及图谱扩展相关工作，也没有直接对知识图谱的实体提取的有效性进行评估。

相比上述研究人员的工作，我们的工作通过引入深入的专利发明点提取机制，实现了专利内容在语义层面上的深度解析和知识点的深度挖掘，不仅在专利数量上达到了较大规模，并且通过高精度的种子知识图谱，加上严谨的图谱扩展方式，对句法规则的有效性进行了实验评估，同时对专利图谱进行了扩展，保障了大规模图谱的质量。虽然本文工作仅仅构建了中医药领域的专利图谱，但是从方法论角度来看，本文的方法具备专利全领域的知识图谱构建技术体系。

3　专利文献种子知识图谱及构建方法

种子知识图谱是智能化知识图谱构建的基础，其中的实体节点多取自于人工搜集或总结的知识，它是智能化知识图谱准确构建的基本保障。相比没有使用种子知识图谱构建的知识图谱，使用了种子知识的图谱构建方法的可靠性相对较高。

3.1　专利种子知识图谱

知识获取的目的是根据非结构化文本构建知识图谱、补全已有的知识图谱以及发现和识别实体和关系[9]。知识获取的任务主要包括关系抽取[10]、知识补全[11]以及实体识别[12]和实体对齐[13]等。

接下来，将以中医药领域为例，分别描述这个领域内种子知识图谱的实体结点的含义、实体间可能存在的几种关系以及种子知识图谱对专利检索、科技创新的作用。

3.1.1　实体节点

中医药领域中的实体结点包含药物/保健品名称、药物成分名称、制作步骤（活动）以及功效名称。这些命名实体也是专利审查人员最关心的技术特征信息，准确捕捉这些信息，是专利审查和分类业务的核心。表 1 用【X】符号标出了专利书《一种玫瑰八宝茶及其制备方法》的标题、权利要求和摘要中的实体结点，其中 X 表示实体结点的名称。

表 1　专利样例节选

Table 1　Excerpt of a patent

标题： 一种【玫瑰八宝茶】及其制备方法
权利要求： 一种【玫瑰八宝茶】，其特征在于其组成原料的重量份为：【玫瑰花】3～5 g、【黄山贡菊】5～8 g、【黄山绿茶】3～5 g、【三七】3～8 g、【决明子】3～5 g、【陈皮】10～15 g、【枸杞】5～10 g、【冰糖】10～20 g。
摘要： 本发明公开了一种【玫瑰八宝茶】及其制备方法，其主要是将组成原料【玫瑰花】、【黄山贡菊】、【黄山绿茶】、【三七】、【决明子】、【陈皮】、【枸杞】和【冰糖】按一定重量份依次通过【烘干】、【粉碎处理】，最后【装袋】得到。本发明制备方法简单，制得的玫瑰八宝茶选用多种上等原料，经过烘干、筛选等处理精制而成，清爽可口、气味芳香，而且具有增强【人体免疫力】的功效。

3.1.2　实体间关系

结合 3.1.1 节给出的案例以及其他未列出的案例，我们发现中医药类专利文献主要涉及以下几种实体间关系。

（1）**上下位关系**[14]：

1）一种【玫瑰八宝茶】及其制备方法；

2）一种治疗肿瘤的【中药组合物】；

3）一种治疗功血的【药剂】。

（2）**整体部分关系中的整体成分关系**：多为中医药配方。下文列举了几种常见的整体成分关系表示，并附上了实例。

1）【玫瑰八宝茶】，其特征在于其组成原料的重量份为：【玫瑰花】3～5 g、【黄山贡菊】5～8 g、【黄山绿茶】3～5 g、【三七】3～8 g……

2）治疗肿瘤的【中药组合物】，其特征在于它包含下列重量份的物质：【牛黄】15～20 份、【白芨】10～30 份、【川贝】10～20 份、【太子参】10～30 份……

3）治疗功血的【药剂】，其特征在于它是由下述重量配比的物质制成：【人面子】8 份、【红豆】9 份、【水芹】9 份、【蔷薇花】5 份、【蓑草】8 份……

（3）**整体部分关系中的阶段活动关系**：多为中医药加工过程。

【玫瑰八宝茶】主要是将组成原料按一定重量份依次通过【烘干】、【粉碎处理】，最后【装袋】得到。

（4）**物体功效关系**：多为中医药的治疗效果。

【玫瑰八宝茶】具有【增强人体免疫力】的功效。

3.2　专利种子知识图谱的构建方法

专利种子知识图谱的构建主要分为 3 个步骤，图 1 给出了专利种子知识图谱的构建流程。下面结合中医药领域专利具体说明专利种子知识图谱的构建细节。

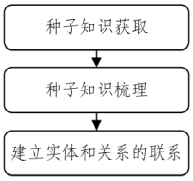

图 1　种子知识图谱的构建流程

Fig. 1　Construction process of seed knowledge graph

3.2.1　种子实体获取

通过与国家知识产权局合作，我们从不同领域的审查组收集到了专利领域专家总结的特征术语。这些术语很多是专利审核员在日常专利审查中的检索要素，是其对相关申请进行阅读后，人工提炼出来的能够代表此专利的技术特征。

不同领域收集到的特征术语表示类型不同、含义相差很大。以中医药领域为例，根据多年人工总结和累积，现已有比较完善的中药名称及其异名信息。在此领域收集到的特征术语被视为中药正名的共有 2963 个，还有一些与正名同义的中药异名，其示例如表 2 所列。

表 2　中药示例

Table 2　Samples of traditional Chinese medicine

编号	中药正名	中药异名
1	三七	田七、山漆、田三七、三七叶、血参、参三七、田漆、金不换
2	肉桂	桂木、桂树、木桂、桂楠、梫、大叶清化桂
3	山药	山芋、野牛尾尻、白苕、薯药、淮山药、儿草、理毛条、延草、玉延
4	丁香	丁子香、雄丁香、公丁香、支解香

3.2.2　种子知识梳理

以中医药领域为例，虽然我们获得了大量的中医药领域的术语，可以把这些术语汇总后作为整体-成分关系中整体

概念和成分概念的命名实体库，但由于专利文献还涉及其他关系类型，如技术领域涉及上下位关系，技术方案除整体成分关系还涉及特征活动关系，技术效果涉及物体功效关系等，因此我们还需进一步总结一些其他概念的命名实体库。表3列出了我们结合中医药领域专利文献以及与专利专业人员合作整理出的中医药领域中文专利涉及的实体关系和实体类别。

表3 中医药领域中文专利涉及的实体关系及实体类别

Table 3 Entity relationship and entity categories involved in Chinese patent on Chinese medicine

关系类别名称	实体类别名称	实体概念举例	备注
上下位关系	上位概念	中药制剂、中药汤剂、中药组合物、装置、设备、结构、制备方法	
	下位概念	《一种预防糖尿病的中药制剂及其制备方法》	
整体成分关系	整体概念	本发明(指代《一种分离中药香精油中残留水的装置》)	
	成分概念	土茯苓、全蝎、制何首乌、杜仲、枳壳、当归、川牛膝、传感器、螺纹柱	
	阶段概念	本方法(指代《一种治疗肝硬化的中药贴剂》)	
	活动概念	烘干，粉碎处理，装袋，加热焖煮，混合，小火翻炒，煎煮	
物体功效关系	物体概念	本发明(指代《一种从丹参茎叶中制备丹酚酸B和丹参素的方法》)	
	功效概念	清热解毒、固表止汗、化湿、交通心肾	
别称关系	—	肉桂/木桂、罗汉果/假苦瓜、三七/田七	此关系下无对应的特定实体类别，“/”前为中文正名，其后为中文异名
因果关系	原因概念	用药量不足，药材生长期短	
	结果概念	中气不足，大气下陷，胁痛	

注：其中整体概念、阶段概念、物体概念作为专利名称的实体链接[15]，指代当前专利

3.2.3 建立实体和关系的联系

经过前面的种子知识梳理过程，我们为种子知识图谱构建做好了准备工作。由于我们构建知识图谱的主要目的之一是协助专利文献检索，因此在构建种子知识图谱的过程中，本体为专利文献，并用业务主键申请号表示。其中的边代表在技术领域、技术方案、技术效果中所涉及的几种实体关系，如上下位关系、整体成分关系、阶段活动关系、物体功效关系；结点则代表成分概念、活动概念、功效概念等实体概念。

经过上述步骤后，我们创建了一个含有194 53个节点、194 775条关系的中医药领域的种子知识图谱。

图2截选了中医药领域的两篇专利文献的种子知识图谱作为展示。

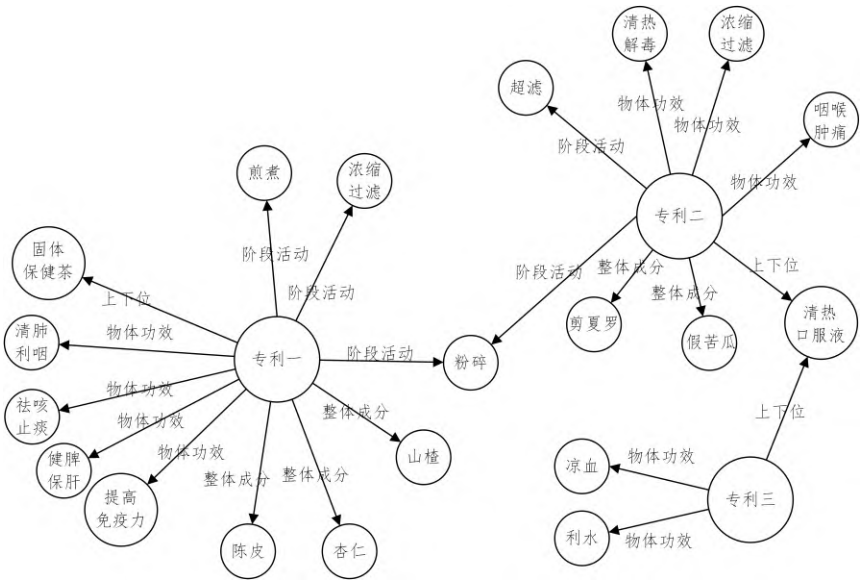

图2 种子知识图谱示例

Fig. 2 Example of seed knowledge graph

4 基于词法句法模式的种子知识图谱扩展方法

前文所述的种子知识图谱的构建可以实现一个小规模的中医药领域知识图谱。由于种子概念的数量相对较小，若要覆盖到整个中医药领域，还需要进行知识图谱扩展。本节通过两种方法对种子知识图谱进行扩展：词法句法模式的人工方法以及自动学习方法。

4.1 词法句法模式的人工总结

通过阅读大量专利，我们发现专利文献中摘要和权利要求部分的结构比较标准和通用，且内容主题相对固定(即技术领域、技术方案与技术效果)，句式和词语搭配也相对固定，对于这种具有固定词语搭配且仅变化其中部分词或短语的语言形式，我们将其称作词法句法模式，如“柴胡具有止咳功效”“人参具有益气养血的功效”这两句话中的词法句法模式可归纳为“…具有…功效”。通过人工阅读专利文献，可以归纳总结出一些句法模式，具体的总结流程如图3所示。4.3节将介绍从文本中自动获取词法句法模式的具体流程。

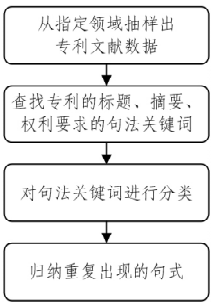

图 3　词法句法模式的总结流程

Fig. 3　Summary flow of lexical and syntactic patterns

由于在词法句法模式中存在大量的相同常量项，为了方便列举词法句法模式，本文定义了其中的常量项，如图 4 所示。

```
Def constant 常量项
{
  标点：，|。|.|？|！|；|：|,|.|？|；|;
  数字：〈0|1|2|3|4|5|6|7|8|9〉*〈.|NULL〉〈0|1|2|3|4|5|6|7|8|9〉*
  量词：份|味|克|两|钱|斤|g|g|枚|片|根|条|只|杯|小杯|对|角|棵|株|
        段|少许|包|个
  是动词：是|作为|成为|为
  有动词：含|有|具有|具备|含有|存在
  能动词：能|能够|可以
  实现词：实现|达到|得到
  方案词：解决|治疗|根治|预防|避免|主治
  效果词：效果|疗效|功效|结果
  对象词：将|由|以
  从属词：属于|从属于
  公开词：公开了|涉及|提供了
  原料词：原料|材料|成分
  包含词：包含|含有|蕴含|包括
  组成词：配置|构成|组成|部件|零件|组件|配件|构件|器件|器材|元件|
          附件|组份
  步骤词：活动|周期|过程|步骤|环节|阶段|操作|程度|单元|动作|方面|
          方式|工序
  后果词：导致|造成|致使|造成
  目的词：为了|是为了
  原因词：因为|因|由于
  存在词：在|位于
  位置词：上面|下面|左面|右面|里面|外面
  下述词：如下|以下|下面|下列
  并列词：和|与|同|及
  关联词：有关|有关系|有联系|相关|相关联
  …
}
```

图 4　词法句法模式中的常量项

Fig. 4　Constant item in lexical and syntactic patterns

下面我们分不同种类列出部分人工总结的词法句法模式。上下位关系句法模式如表 4 所列，整体成分关系句法模式如表 5 所列，阶段活动关系句法模式如表 6 所列，物体功效关系句法模式如表 7 所列，其他关系中的句法模式如表 8 所列。

表 4　上下位关系句法模式

Table 4　Syntactic patterns of hyponymy relationship

句法模式	举例	说明
一种【?上位概念】	一种脑梗塞的中药制剂	
【?下位概念】〈公开词〉一种【?上位概念】	本发明提供一种止痒联合用药物	
【?下位概念】〈是动词〉一种【?上位概念】	本药物发明为一种鱼鳞胶的制备方法及其产品	
【?下位概念】〈从属词〉【?上位概念】领域	本发明属于药物制剂领域	
【?下位概念】等【?上位概念】	三七、枸杞、半枝莲等药用植物	len(下位概念)>1
【?上位概念】〈包含词〉【?下位概念】	钙离子通道阻断剂类药物还包括硝苯地平、维拉帕米等	
【?上位概念】〈有动词〉【?下位概念】等	保肝类中草药有五味子、当归、三七、丹参等	len(下位概念)>1
【?上位概念】〈有动词〉【?下位概念】几种	目前，治疗高血压常用的交感神经抑制类药物有可乐定、利美尼定、利血平、普萘洛尔几种	len(下位概念)>1

表 5　整体成分关系句法模式

Table 5　Syntactic patterns in overall-component relationship

句法模式	举例	说明
【?整体概念】〈对象词〉【?成分概念】为	一种缓泻通便的中药丸剂，它以麻子仁、杏仁、芍药、厚朴、枳实为原料煎制成	本案例在后面会多次用到
【?整体概念】〈包含词〉【?成分概念】	该洗面奶包括 16～18 混合醇、硬脂酸、棕榈酸异辛脂、氢氧化钾、维生素和去离子水	len(成分概念)>1
【?整体概念】〈对象词〉【?成分概念】〈组成词〉	中药养心汤由人参、丹参、灵芝、当归、肉桂、桂皮、黄芪、冬麦、甘草、红花、芦荟、葛根组成	
【?整体概念】〈组成词〉〈原料词〉〈是动词〉:【?成分概念】	本玫瑰花粉豆浆组成原料为：黑豆粉、莲子粉、蚕豆粉、玫瑰花粉和人参	len(成分概念)>1
【?整体概念】所〈包含词〉的【?成分概念】	其特征在于该 EVA 胶膜所包含的交联固化剂、成核增透剂、光稳定剂和增粘剂按照一定比例配制	
【?整体概念】〈有动词〉【?成分概念】等	本发明中组合物有黑木耳提取物、红曲提取物、山楂提取物、蒲黄提取物等多种具有降血脂功能的生物活性物质	len(成分概念)>1
【?成分概念】〈组成词〉了【?整体概念】	本发明是以铑化合物作为羰基合成的催化剂，以烃基碘为助催化剂，辅以季胺盐促进剂，以锌化合物为稳定剂，组成了羰化法制备乙酸的催化体系	
【?成分概念】〈数字〉〈量词〉	本发明涉及一种保健酒，由下列重量份的原料制成：金银花 11 克、生地黄 11 克、人参 11 克	

表 6　阶段活动关系句法模式

Table 6　Syntactic patterns in phase-activity relationship

句法模式	举例	说明
【?阶段概念】通过【?活动概念】〈实现词〉	本制备方法通过烘干、粉碎处理，最后装袋得到中药材成品	len(活动概念)>1
【?阶段概念】〈包含词〉【?活动概念】等	该制备过程包括烘干、粉碎、混匀、杀菌及烘干打包等步骤	len(活动概念)>1
【?阶段概念】〈包含词〉〈下述词〉〈步骤词〉:【?活动概念】	所述的人参果软胶囊的制备方法包括以下步骤：1）人参果甙的制备；2）刺五加提取物的制备；3）提取物混合后超微粉碎；4）压制成软胶囊	len(活动概念)>1

（续表）

句法模式	举例	说明
【?阶段概念】〈对象词〉〈下述词〉〈步骤词〉〈组成词〉:【?活动概念】	本制备方法由下列步骤组成:初步乳化,酶解交联,乳化反应,洗涤精制,均质脱色,干燥包装	len(活动概念)>1
【?阶段概念】的〈步骤词〉〈包含词〉【?活动概念】	本发明的步骤包含合成中间体喹噁啉-1,将该中间体在甲醇和浓盐酸中进行醇解,与氰乙酰肼反应合成最终产物	len(活动概念)>1
【?阶段概念】的〈步骤词〉〈包含词〉〈下述词〉:【?活动概念】	本培育方法的步骤为:1)制备培养基;2)菌种制作;3)蛹体接菌;4)鲜虫草成型	len(活动概念)>1
【?阶段概念】的〈步骤词〉〈有动词〉【?活动概念】	该方法包括的步骤为:1)获取无菌苗;2)诱导培养;3)增殖培养;4)生根培养;5)营养培护	len(活动概念)>1
【?阶段概念】的〈步骤词〉〈有动词〉〈下述词〉:【?活动概念】	本中药制剂的配制步骤为:1)除杂;2)筛选烘干;3)振动;4)清理	len(活动概念)>1

表 7 物体功效关系句法模式

Table 7 Syntactic patterns in object-effect relationship

句法模式	举例
【?物体概念】〈有动词〉【?功效概念】〈效果词〉	本发明中药组合物具有养血柔肝、补肺健脾、益肾生精、扶正固本的功效
【?物体概念】〈有动词〉〈下述词〉〈效果词〉:【?功效概念】(len(功效概念)>1)	本口服液具有以下功效:清热解毒、凉血散瘀、益气生津、调和脾胃、消肿散结
【?物体概念】〈实现词〉【?功效概念】〈效果词〉	本发明能实现拔罐吸附于人体时同时进行针灸与释放药物的效果
【?物体概念】〈能动词〉【?功效概念】	制得的护手霜能够有效缓解疲劳,促进血液循环,达到暖肤的功效
【?物体概念】〈能动词〉〈实现词〉【?功效概念】〈效果词〉	并且传递的活性成分能够达到一定的靶向治疗效果
【?物体概念】〈方案词〉【?功效概念】	所述抗体可用于预防、治疗和诊断破伤风感染和/或治疗破伤风梭菌感染介导的一种或多种症状
【?物体概念】的〈效果词〉〈包含词〉【?功效概念】	本药品的功效包括消除或减轻口臭,除渍增白,牙齿结石,预防出血,预防蛀牙等功效
【?物体概念】〈包含词〉〈下述词〉〈效果词〉:【?功效概念】(len(功效概念)>1)	去乙酰毛花苷包含以下功效:纠正心衰、纠正心率失常

表 8 其他关系中的句法模式

Table 8 Syntactic patterns in other relationships

句法模式	举例	说明
【?主体概念 1】,又称【?主体概念 2】	罗汉果,又称假苦瓜	体现一种别称关系
【?药物概念】〈方案词〉【?疾病概念】	中药制剂根治气管炎	体现一种药物治疗关系
【?药物概念】〈能动词〉〈方案词〉【?疾病概念】	此药膏能够根治外痔肛裂	体现一种药物治疗关系
【?原因概念】〈后果词〉【?结果概念】	压迫时间不足所导致的皮下渗血的技术问题	体现一种因果关系
【?结果概念】〈原因词〉【?原因概念】	这种病情的恶化因为晚换药而造成	体现一种因果关系
【?主体概念 1】〈并列词〉【?主体概念 2】〈关联词〉	朊病毒和牛海绵状脑病(也称为疯牛病)有关系	体现一种关联关系
【?主体概念 1】〈并列词〉【?主体概念 2】之间〈有动词〉〈关联词〉	前述的高血压、高血脂病症与人体的心、脑、肾之间有着密切的联系	体现一种关联关系
【?部分概念】〈存在词〉【?主体概念】的〈位置词〉	腿部模块在腰背模块的下面	体现一种位置关系
【?主体概念】的〈位置词〉〈是动词〉【?部分概念】	床的下面是按摩器、振动器、拍打器和红外线发生器	体现一种位置关系

注:带有“?”符号的部分都是需要提取的技术特征词,如:【?主体概念】和【?部分概念】;带有“〈〉”符号的部分代表之前定义的不同类别的常量词,如〈方案词〉

表 8 中,有些词法句法模式加入了条件限制,可以使得该词法句法模式匹配到的概念更加精准。例如下面的阶段活动词法句法模式:

【?阶段概念】〈包含词〉【?活动概念】(len(活动概念)>1)

如果加上限制条件(len(活动概念)>1),就可以匹配到包含正确概念的句子,如“本中药配方过程包含选材、称量、混合和煎制等步骤”,其中“选材”“称量”“混合”“煎制”都是正确的活动概念。

如果不加上述限制条件,则可能出现错误匹配,如“本中药配方过程包含多个简单加工步骤”,其中“多个简单加工”不是一个合理的活动概念。

4.2 词法句法模式匹配专利文献的算法流程

将 4.1 节中通过人工总结的词法句法模式应用到中医药领域专利文献中需要相关的算法流程。词法句法模式的匹配流程如算法 1 所示。

算法 1 词法句法模式匹配专利文献的算法流程

1. 对中医药领域专利文献中的标题、摘要和权利要求进行预处理(包括文本内容分词和词性标注),为词法句法模式匹配做准备。
2. 将〈上下位关系〉、〈整体成分〉、〈特征活动〉、〈物体功效〉词法句法模式应用在标题、摘要和权利要求中匹配出技术领域特征。
3. 将提取到的新的技术特征词汇进行人工审核,将符合要求的加入到先前收集的种子概念词典中进行扩充。需要指出的是,这种靠词法句法模式发现的技术特征词汇并不是百分之百都正确,因此通过人工审核可以保证技术特征发现的质量。

算法说明:

(1)上下位关系句法模式可以匹配得到以下结果:

一种【玫瑰八宝茶】及其制备方法

【本发明】公开了一种【奇蒿总黄酮】的制备方法

【本发明】涉及一种治疗风湿性关节炎的【中药汤剂】

(2)整体成分关系句法模式可以匹配得到以下结果:

【玫瑰八宝茶】组成原料为:【玫瑰花】、【黄山贡菊】、【黄山绿茶】…

下述重量配比的物质制成:【人面子】8 份、【红豆】9 份、【水芹】9 份…

(3)特征活动关系句法模式可以匹配得到以下结果:

依次通过【烘干】、【粉碎处理】,最后【装袋】得到

(4)整体成分关系句法模式可以匹配得到以下结果：

具有【增强人体免疫力】的功效

达到【有效麻醉】的显著功效

有【增强免疫】的作用

4.3　词法句法模式的自动学习和知识图谱的自动扩展

由于人工句法模式的总结过于耗时，且不能涵盖全部的词法句法模式。本节将着重讨论如何利用已经获取的种子知识图谱进行扩展。

利用人工归纳总结的种子概念，我们提出了一种基于种子概念词法句法模式的自动学习方法；并通过新学习的词法句法模式的进一步迭代，得到了新的命名实体。词法句法模式的自动学习和知识图谱的自动扩展的过程如算法 2 所示。

算法 2　词法句法模式的自动学习和知识图谱的自动扩展

1. 领域词典构造：构造中医药领域的种子概念词典。
2. 词法句法模式学习数据集构造：在中医药领域专利文献中，抽取带有指定关系的语句，形成词法句法模式学习数据集。
3. 词法句法模式的自动学习：基于上述词法句法模式学习数据集进行学习，并且汇总置信度较高的优质词法句法模式，作为自动学习出来的扩展词法句法模式。
4. 新知识点获取：将学习出来的扩展词法句法模式应用到更多中医药领域的文献中，从而发现更多的关系对实例，对关系对进行筛选，保留正确的关系对，为知识图谱的扩展准备数据。
5. 知识图谱扩展：将上面获取到的知识点（包括实体概念和关系实例）扩充到知识图谱中，需要针对新知识点和原知识图谱的关系，考虑多种不同情况下的扩充方式。

下面将对以上几个步骤进行详细说明。由于专利文本中关系种类涵盖较多，如上下位关系、整体成分关系、因果关系等，因此以下内容仅以中医药领域的整体成分关系进行举例说明。

4.3.1　中医药领域种子概念词典

整体成分关系是语言学的概念之一，其中组成要素属于成分概念，而多种组成要素构成的整体属于整体概念。在中医药领域的专利文本中，整体概念一般指中药、药剂等成型的药物处方，而成分概念指不同的中医药材料，用于构成不同的药物。

我们通过与知识产权局相关领域的审查组专家进行合作，并参照中医药学语言系统的语义网络框架[16]，构造了一定规模的种子概念词典，包括中医药专利整体概念词典和中医药专利成分概念词典，部分示例如表 9 所列。需要说明的是，本节中的整体概念和成分概念包含在 3.2.3 节的种子知识图谱中。

表 9　中医药专利种子概念词典示例

Table 9　Examples of Chinese medicine patent seed concept dictionary

概念名称	概念示例
中医药专利整体概念	本专利（指代《一种预防糖尿病的中药制剂及其制备方法》） 本发明（指代《一种分离中药香精油中残留水的装置》）
中医药专利成分概念	三七、丝瓜络、丹参、乳香、人参、党参、全蝎、制何首乌、制川乌、制草乌、土茯苓、地黄、地龙、大黄、天麻、威灵仙、山茱萸、山药、川牛膝、川芎、巴戟天、当归、忍冬藤、木通、杜仲、枸杞、柴胡、栀子、桃仁、桑寄生、桔梗、款冬花、蜈蚣、壁虎、传感器、螺纹柱

4.3.2　词法句法模式学习数据集构造

词法句法模式需要基于一定量的中医药领域的相关数据集进行学习。学习数据集要有足够量并且包含需要发现的中医药领域整体概念词和成分概念词。在中医药专利领域中存在大量的整体概念和成分概念被一个标点符号隔开的情况，如："一种缓泻通便的中药丸剂，它以麻子仁、杏仁、芍药、厚朴、枳实为原料煎制成"。其中整体概念为"一种缓泻通便的中药丸剂"，和后面的成分概念被逗号隔开，并且通过代词指代。因此，我们在构造学习数据集时，需要将以句号为结尾的整个句子作为一条数据。下面是数据集构造的过程。

第一步，选择 10000 篇医疗领域专利文献的权利要求和摘要进行分句子和分词工作；第二步，对于每一个句子，基于前面构造的种子概念词典，标识出概念词汇，汇总所有同时包含整体概念和成分概念的语句，共计 9548 句；第三步，将句子中整体概念和成分概念词分别用〈@整体概念〉和〈@成分概念〉标签进行替代，将以并列形式出现的多个同类概念合并成一个标签。

上述过程中的最后一步是为给后面的词法句法模式学习做准备。以前面提到的语句为例：一种缓泻通便的中药丸剂，它以麻子仁、杏仁、芍药、厚朴、枳实为原料煎制成。通过标签替代和合并得到下面的结果（句子中的"/"符号表示分词结果）：一种/缓泻通便/的/〈@整体概念〉，它/以/〈@成分概念〉/为/原料/煎制/成。

4.3.3　句法模式自动学习

4.3.2 节整理出了 9548 条带有整体和成分概念的语句。针对这些语句，采用统计和聚类等算法挖掘其中的高频重复模式，形成学习到的词法句法模式（注：词法句法模式学习指根据已知的整体概念和成分概念，学习到频繁出现的前缀、后缀和中间项）。学习过程如算法 3 所示。

算法 3　自动学习词法句法模式的详细流程

1. 构造词法句法模式的基本形式：

〈?前缀〉〈@整体概念〉〈?中间项〉〈@成分概念〉〈?后缀〉或

〈?前缀〉〈@成分概念〉〈?中间项〉〈@整体概念〉〈?后缀〉

2. 对学习的数据集中所有句子的形式进行统计，获取重复度超过一定阈值的句子形式作为候选句法模式，本次工作取阈值大于或等于 3，意味着只要同一种模式出现了 3 次以上（含 3 次），就可以作为一个候选词法句法模式，这样统计出来的词法句法模式达到了 800 多个；
3. 根据找到的整体成分关系对对上面统计出来的词法句法模式进行聚类，聚类相似度阈值 $\lambda=0.5$，本次聚类针对 800 多个词法句法模式聚类出 195 类效果相似的词法句法模式类；
4. 对聚类后的每一类中的所有候选词法句法模式在数据集上出现的次数进行求和，作为该词法句法模式类中的词法句法模式的关系命中条数。取命中条数中 top-10 类词法句法模式作为学习到的新的句法模式，其所包含的所有句法模式有 53 条。

以上步骤中有如下几个关键点需要进一步解释。

第一点：其中句子中的前缀和后缀只选取紧挨概念词的左边或者右边的第一个词，如果第一个词是助词，则删除不要。例如："一种/缓泻通便/的/〈@整体概念〉，它/以/〈@成分概念〉/为/原料/煎制/成"。经过处理，它的句法模式为："〈@整体概念〉，它/以/〈@成分概念〉/为"。其中前缀部分为整体概念左边第一个词"的"，作为助词连接前面的定语，因此

删除不要前缀项，只保留后缀和中间项。这样做的好处是保留了句式中的固定部分，忽略了定语、状语等可变部分，有助于提取通用词法句法模式。

第二点：词法句法模式聚类以找到的整体成分关系实例为依据判断相似性。所谓关系实例就是一对整体成分概念组，例如：(中药汤剂，当归)以及(中药方，三七)等。我们采用Jaccard算法来评价两个词法句法模式的相似度，原理如下：

(1)词法句法模式 a 匹配的关系实例数为 $R(a)$，词法句法模式 b 匹配的关系实例数为 $R(b)$。

(2)词法句法模式 a 和词法句法模式 b 匹配的共同实例关系数为：$R(a)\cap R(b)$。

(3)词法句法模式 a 和词法句法模式 b 匹配的总共实例关系数为：$R(a)\cup R(b)$。

(4)则 A 和 B 的相似度由下面的公式定义：

$$Sim(a,b)=|R(a)\cap R(b)|/|R(a)\cup R(b)|$$

第三点：前文定义了两个词法句法模式 a 和 b 之间的相似度。对于两个词法句法模式类来说，其相似度等于它们之间任意两个词法句法模式的相似度的平均值。其计算方式如下：

(1)词法句法模式类 A 中包含词法句法模式 $\{a_1,a_2,\cdots,a_m\}$。

(2) 词法句法模式类 B 中包含词法句法模式 $\{b_1,b_2,\cdots,b_n\}$。

(3) 则类 A 和类 B 的相似度为：

$$Sim(A,B)=\Sigma_i\Sigma_j Sim(a_i,b_j)/(m*n)$$

第四点：基于上面的词法句法模式和词法句法模式类的相似度定义，本文采用下面的层次聚类方法来实现词法句法模式类的聚类，流程为：第一步，将所有学习到的词法句法模式 $r_1\ r_n$，转化为单元素的词法句法模式类 $\{r_1\}\{r_n\}$；第二步，根据定义的相似度公式求出两两词法句法模式类的相似度值 $Sim(\{r_i\},\{r_j\})$；第三步，找到相似度值最大的两个词法句法模式类，若它们之间的相似度值大于一个阈值 $\lambda=0.5$，则合并这两个词法句法模式类，即 $\{r_i\}+\{r_j\}=\{r_i,r_j\}$，否则本层次聚类算法结束；第四步，求出在第三步中新生成的词法句法模式类和其他类的相似度值，跳转到第三步做循环。

表10列举了一些自动学习出来的词法句法模式，可以看出，学习出来的词法句法模式具备一定的准确性，证明了此方法的有效性。

表10 自动学习到的词法句法模式示例

Table 10 Examples of automatically learned syntactic patterns

学习到的词法句法模式	匹配的例句
〈@整体概念〉，它以〈@成分概念〉为	一种缓泻通便的中药丸剂，它以麻子仁、杏仁、芍药、厚朴、枳实为原料煎制成
〈@整体概念〉，它是由〈@成分概念〉为	一种治疗痔疮的药物，它是由炙黄芪、炙甘草、炒山药和升麻等为原料制成
〈@整体概念〉是由〈@成分概念〉按	所述的中药组合物是由重楼皂苷提取物和丹参醇提物按1:1～1:4配伍所组成
〈@整体概念〉，主要由〈@成分概念〉组成	一种治疗疝气的中药组合物，主要由赤茯苓、槟榔、陈皮、山木香、川楝子、当归组成
〈@整体概念〉，其重要组分为：〈@成分概念〉组成	一种治疗慢性荨麻疹的药及其制备方法，其重要组分为：当归、生地、桃仁、蝉退、防风、丹参、赤芍、荆芥、桑叶、紫草、白蒺藜、甘草组成

4.3.4 新实体和新关系获取

通过上文学习到的新的词法句法模式，采用模式匹配的方法对大量的中医药领域专利文献的权利要求和摘要部分进行整体概念和成分概念的抽取。概念抽取的过程和词法句法模式抽取的过程类似，但关注点不同。在词法句法模式抽取时，通过已知的概念来找到词法句法模式中的前缀、后缀和中间项；在新知识点获取时，通过已知的前缀、后缀和中间项来发现整体概念和成分概念。

抽取过程为：第一步，对每一篇待分析的中医药领域专利文献的权利要求和摘要进行分句处理；第二步，应用每一条学习到的词法句法模式对每一个句子进行概念抽取。

在知识抽取过程中，我们遵循如下的词法句法模式：

(1)如果词法句法模式包含前缀、后缀和中间项，那么前缀和中间项之间部分为整体概念，中间项和后缀之间部分为成分概念。需要注意的是，成分概念大多是由多个并列概念组成。

(2)如果句法模式缺乏前缀，那么句子开始到中间项部分都是整体概念或者成分概念。

(3)如果句法模式缺乏后缀，那么从中间项到句子末尾部分都是成分或者整体概念。

(4)在实验中，尚未发现缺乏中间项的情况，因此不予考虑。

通过上述抽取过程，我们发现大量的原始概念(词或者短语)和相关关系实例还需要经过以下几个步骤进行深入处理，才能认定所发现的新关系实例是正确的。具体流程如算法4所示。

算法4 判定新关系实例是否正确的流程

1. 去冗余处理：对每一个关系实例中出现的原始概念进行特征词提取，剥离冗余部分，形成特征概念。
2. 去重处理：对所有特征概念关系实例进行查重处理，相同的特征概念关系实例只保留一份，并且记录重复的数字，如Count(玫瑰八宝茶，玫瑰花)=5。
3. 去除已存在的关系实例：将去重后的整体成分关系实例和种子知识图谱中的关系实例进行比对，去掉已经在种子图谱中出现过的关系。
4. 验证新关系：根据并列关系以及重复次数对剩下的新关系实例进行验证，确认置信度高的新关系。
5. 人工验证：最终剩下的新关系无法自动验证，需由人工进行验证。

其中，去冗余处理是将混杂了其他词语的整体或者部分概念词语进行还原，只保留特征词部分，也就是概念词。这种情况只会出现在缺少前缀或者后缀的句法模式匹配中。具体示例如表11所列。

表11 句法模式示例

Table 11 Example of syntactic pattern

学习到的句法模式	匹配的例句
〈@整体概念〉，它以〈@成分概念〉为	一种缓泻通便的中药丸剂，它以麻子仁、杏仁、芍药、厚朴、枳实为原料煎制成

表11中左边的句法模式可以匹配到右边例句。由于句法模式具有后缀，因此成分概念部分在中间项和后缀之间非常清晰，而整体概念部分是从句首到中间项部分：

一种缓泻通便的中药丸剂

这样提取出来的整体概念显然不是真正的整体概念，其中包括了不需要的冗余部分“一种缓泻通便的”，需要剥离掉。通过分析大量的中医药领域专利文献，我们总结出了3条冗余部分剥离的原则。第一条，首先对原始概念词进行分词处理，并且标注词性；第二条，对于缺少前缀的词法句法模式匹配，选取紧挨着中间项之前的连续名词词组构成的特征词，以及与这个特征词具有并列关系的其他特征词一起作为概念词；第三条，对于缺少后缀的词法句法模式匹配，选取紧挨着中间项之后的连续名词词组构成的特征词，以及和这个特征词具有并列关系的其他特征词一起作为概念词。

下面以缺少前缀为例说明如何剥离冗余部分，以提取真正概念词，如表12所列。

表12　冗余剥离示例

Table 12　Examples of redundancy stripping

原始概念词组	去冗余处理
一种/mq 缓泻通便/v 的/u 中药/n 丸剂/n，它以麻子仁……	根据句法模式，中间项“，它以”之前最近的名词组合为“中药丸剂”，因此它是整体概念词
一种治疗慢性荨麻疹的药及其制备方法，其重要组分为：当归……	根据句法模式，中间项“，其重要组分为：”之前最近的名词组合为“制备方法”；同时通过“及其”和前面的名词“药”发生并列关系，因此“药”和“制备方法”都是本词法句法模式提取的整体概念词

经过去冗余处理后，新发现的关系实例中的整体和部分概念词汇都会比较“干净”，然而它们构成的新关系实例不能直接被认定为是正确的。由于新发现的关系实例众多，若都采取人工认定方式则工作量巨大。通过实验发现，遵循以下原则进行机器处理可以比较准确地筛选出可信的新关系实例。

(1)如果一个关系实例重复出现的次数大于一个阈值，则可以认为是正确的关系实例。本实验采用的阈值为$\lambda=5$。需要指出的是，极个别大量出现的关系实例不一定正确，可以通过添加黑名单词典的方式进行过滤。比如前面案例中“制备方法”这个词不属于真正的整体概念，需要进行黑名单处理。

(2)如果一个关系实例中的成分概念和种子成分概念词典中的某个概念并列出现于同一个整体成分词法句法模式匹配实例中，则可以认为此关系实例是正确的。例如，在前面提到的案例“中药丸剂，它以麻子仁、杏仁、芍药、厚朴、枳实为原料”中，如果新发现的关系是(中药丸剂，厚朴)，而同在此次词法句法模式匹配中发现了一个种子知识图谱中存在的关系(中药丸剂，杏仁)，由于厚朴和杏仁为并列关系，则可以推断出新发现关系(中药丸剂，厚朴)也成立。

通过上述一系列新关系实例认定过程，绝大多数关系实例可以实现自动认定，剩下的无法认定部分需要由人工来进行筛选。筛选之后的所有新关系，都可以扩充到种子知识图谱中，同时，新关系中的整体概念和部分概念词汇如果不在种子概念词典中，也可以扩充到种子概念词典中，为将来的滚动学习奠定基础。

4.3.5　知识图谱扩展方法

4.3.4节中，我们已经获取了新的实体概念和关系实例，这些新知识需要插入到原先的知识图谱中进行扩充。在图谱扩充过程中，根据新的实体概念和关系实例在已有图谱中存在与否可以分成以下几种情况。

第一种情况：新发现的关系实例中两个实体概念都已经存在，但是新的关系不存在。这种情况下，可以借助4.3.4节的方法在图谱中加一条边连接两个实体概念，边的标签为新的关系。例如，图5中(a)标记的区域表示专利一与成分概念“金星草”新建立起“整体成分”关系，(b)标记的区域表示专利二中的成分概念“假苦瓜”与专利三中的结果概念“凉血”和“利水”分别新建立起“物体功效”关系，从而对知识图谱进行了扩展。

第二种情况：新发现的关系实例中一个实体概念存在，另外一个实体概念不存在。例如：图6中(a)标记的区域中实体概念“煎煮”及对应的阶段活动关系和(b)标记的区域中实体概念“苦地丁”及对应的整体成分关系。扩展流程如图6所示，仅有一个实体概念的图谱扩展如算法5所示。

算法5　仅有一个实体概念的图谱扩展

1. 对于新加入的任意新的实体概念节点Y，可首先通过4.3.4节中发现的关系实例找到和Y发生关系的原有节点X，作为Y节点的接入节点，表明Y节点是通过X这个原有节点连通到知识图谱的；
2. 如果找不到接入节点X，则建立Y节点和它所在的专利文献节点P之间的索引关系，表明Y节点是P文献的一个索引词；
3. 找到和接入节点X在图谱中连接半径为三跳的其他所有邻居节点，这些邻居节点构成新关系发现的候选节点集合$S=\{X_0, X_1, \cdots, X_n\}$；
4. 通过计算新加入节点Y和集合S中所有节点的相似度，可以发现新加入节点Y和其他已有节点的关系实例，从而实现知识图谱的补全。

此外，我们通过TransD[17]模型也发现了一些实体间新的关系，如图7中(a)中使用“*”标记的关系，从而对知识图谱进行补全，提高集聚系数。全部情况列举完毕后，将详细介绍TransD的相关内容。

第三种情况：新发现的关系实例中两个实体概念和他们之间的关系实例均不存在，如图7中(a)区域中的两个实体结点“血竭”和“活血化瘀”及其物体功效关系。扩展流程如算法6所示。

算法6　两个实体概念和其间关系均不存在的图谱扩展

1. 先将两个新实体概念扩充到知识图谱中，然后添加一条连接两个实体概念的关系边。
2. 借助同义概念关系，找到与这两个未曾与图谱连通的新加的实体有相关关系的节点，即通过词嵌入(Word Embedding)方法将不同的实体概念转化为向量，然后利用Jaccard方法求出向量的相似度。
3. 当两个词向量的相似度达到一定阈值(如Jaccard(vec1，vec2)＞0.9)时，可认为它们是同义词。
4. 当实体概念被引入后，通过TransD模型将此节点与现有实体间的关系实例进行推理，实现知识图谱补全。

接下来，将介绍如何借助TransD翻译模型对新发现实体和已有实体进行关系补全，如算法7所示。

算法 7 TransD 扩展实体间关系

1. 将人工采集的包含上下位关系、整体部分关系、物体功效关系的语句中的实体概念和关系概念分别拆分出来，构建成实体概念训练集和关系概念训练集。
2. 其次，通过 TransD 模型将各实体和关系概念采用 ADADELTA SGD 优化方法和超参数 margin=1，m=n=50，batch_size=200 组合，并使用 L_2 正则项将各个实体训练成两个向量（表示头实体的两个向量用 h 和 h_p 表示，表示关系的两个向量用 r 和 r_p 表示，表示尾实体的两个向量用 t 和 t_p 表示，其中 $h, h_p, t, t_p \in \mathbb{R}^n, r, r_p \in \mathbb{R}^m$），一个表示其语义信息的 n=50 维的向量和一个用于构建实体到关系的映射关系矩阵的 n=50 维的向量。
3. 构建好映射关系的三元组可以表示为 (h', r, t')，且可以根据 $h'+r=t'$ 的原则进行知识补全。

通过上述图谱扩展方法，我们把 4.3.4 节获取的新知识扩充到原先的知识图谱中，得到扩展后的中医药领域中文专利知识图谱。由局部示例图 5 可知，当审核员检索到专利发明对象“专利一”时，其可通过“粉碎”这一实体节点关联到具有相同节点的专利发明对象“专利二”，进而可通过成分概念“假苦瓜”关联到具有相同功效“利水”和“凉血”的专利发明对象“专利三”。由图 6 可知，当审核员检索到专利发明对象“专利四”时，其可通过“桔梗”这一实体节点关联到具有相同节点的专利发明对象“专利五”。这种图谱中专利发明对象间的相互关联有助于审核员准确、高效地检索相关信息，进而做出准确判断。

图 7 中带有数字标号和虚线的实体和关系实例为新扩充知识，扩充结果示例如表 13 所列。

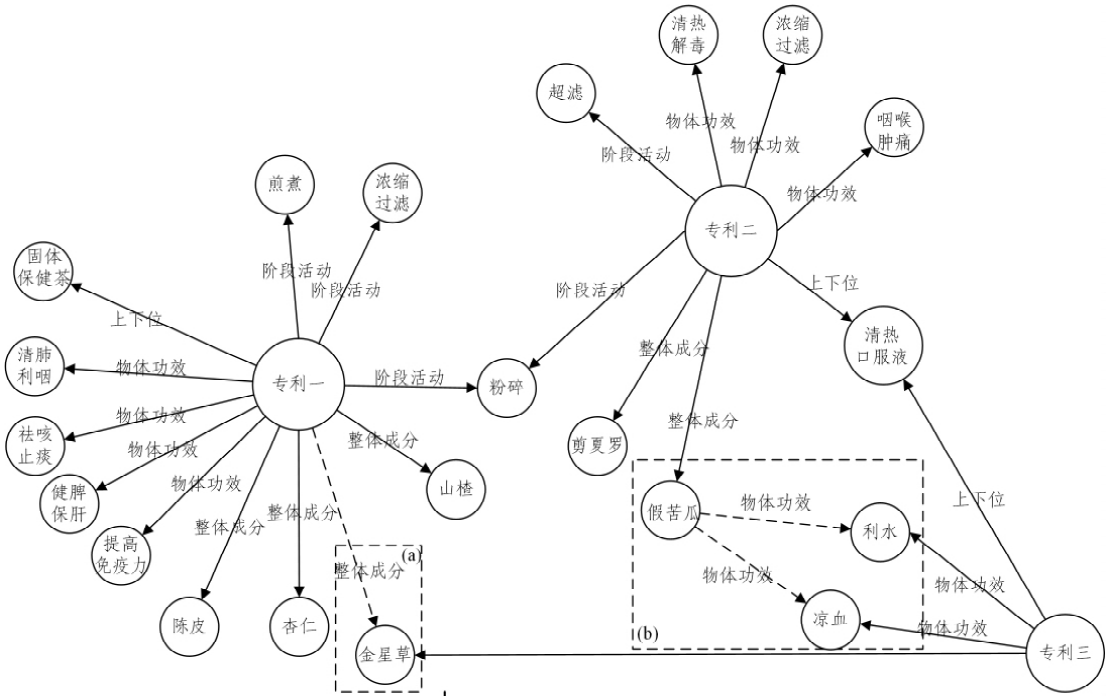

图 5 扩展后的知识图谱示例 1

Fig. 5 Extended knowledge graph example 1

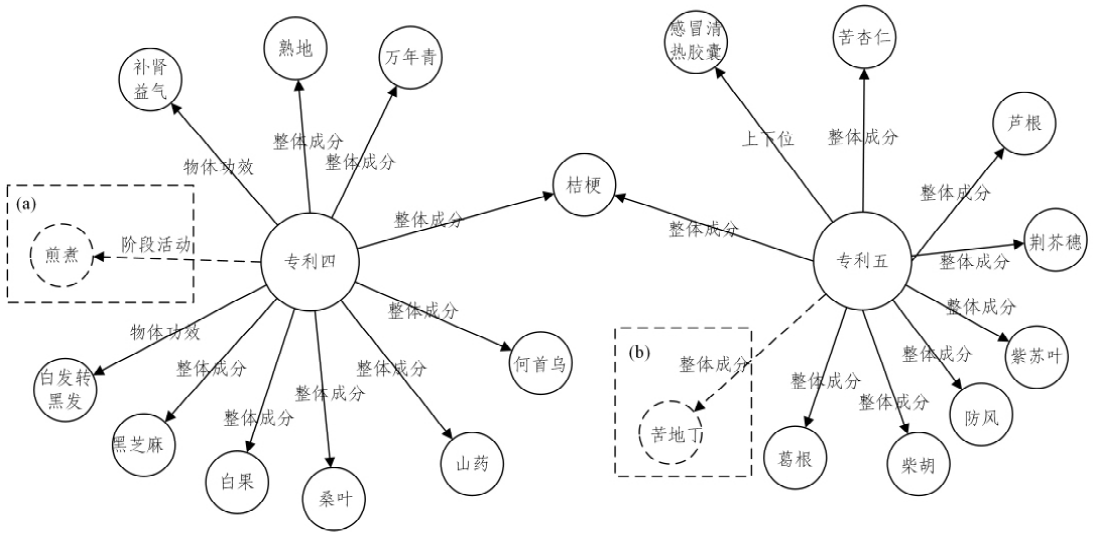

图 6 扩展后的知识图谱示例 2

Fig. 6 Extended knowledge graph example 2

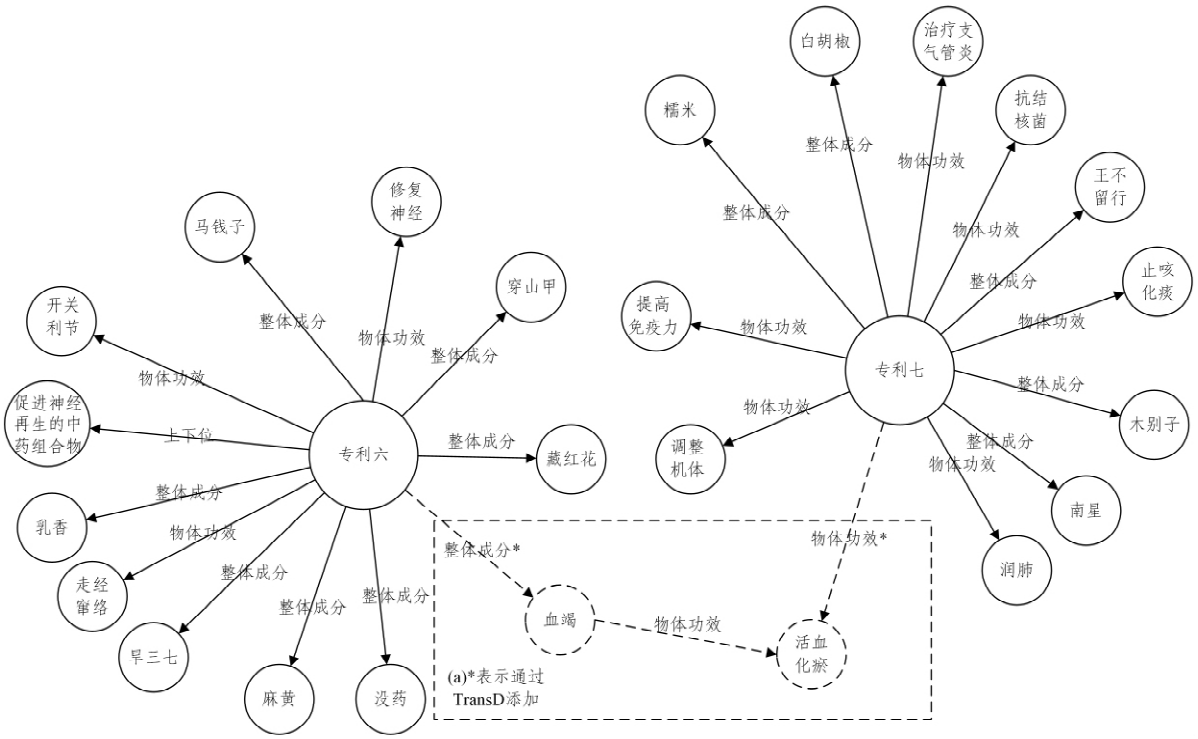

图 7　扩展后的知识图谱示例 3

Fig. 7　Extended knowledge graph example 3

表 13　扩充结果示例

Table 13　Examples of extended results

扩充编号	扩充类别	扩充结果
(1)	实体概念都存在，关系不存在	成分：假苦瓜 功效：凉血，利水 关系：物质功效（新添加）
(2)	1 个实体概念存在，另外一个实体概念不存在	阶段：专利二 活动：烘干（新添加） 关系：阶段活动（新添加）
(3)	两个概念均不存在	物体：金钏草（新添加） 功效：清肺泄热（新添加） 关系：物体功效（新添加）

5　实验结果及分析

在种子知识图谱的基础上，我们使用机器学习方法学习得到新句法模式后，使用 4.3.4 节中介绍的方法对种子知识图谱中实体概念和实体关系分别进行了扩展，其中下位概念节点是人工添加的。同时，我们又使用翻译模型 TransD 对实体关系进行进一步扩展，完成中医药领域中文专利的知识图谱的构建。实验结果汇总如表 14 和表 15 所列。为验证本文中机器学习的句法模式的有效性，我们选择了 1 000 篇中医药专利领域文献作为测试数据，对其中的实体概念做了人工标注。

在测试过程中，我们分别用人工总结的词法句法模式和机器学习的词法句法模式对 1 000 篇专利测试文献进行实体和关系挖掘，并且与标注数据进行比较来判断句法模式的有效性。测试的结果如表 16 所列。

从表 14—表 16 可以看出，不管是人工词法句法模式还是机器学习词法句法模式，都发现了大量的正确的实体节点和实体关系，可以有效提高知识获取的覆盖度，同时扩充了中医药领域知识图谱的规模。目前，人工总结的词法句法模式的准确率高于机器学习的词法句法模式的准确率，未来我们还会不断优化机器学习方法，进一步提升准确率。

表 14　中医药领域中文专利知识图谱构建的实体扩展实验结果

Table 14　Results of entity expansion experiments on construction of Chinese patent knowledge graph in traditional Chinese medicine

实体类别	种子图谱中的实体数量/个	扩展后的数量/个	增长比例/%
上位概念	238	359	50.8
下位概念（专利名称）	10 000	545 233	5 450.0
整体概念	专利指代词	—	—
成分概念	3 233	4 178	29.2
阶段概念	专利指代词	—	—
活动概念	1 129	1 420	25.8
物体概念	专利指代词	—	—
功效概念	502	876	74.5
原因概念	1 113	1 329	19.4
结果概念	3 762	4 448	18.2
合计	19 453	558 461	2 770.0

注：整体概念、阶段概念、物体概念作为实体链接均指代专利名称

表 15　中医药领域中文专利知识图谱构建的关系扩展实验结果

Table 15　Results of entity expansion experiments on construction of Chinese patent knowledge graph in traditional Chinese medicine

关系类别	种子图谱中的数量/个	扩展后的数量/个	增长比例/%
上下位关系	10 000	472 879	4 728.80
整体成分关系	96 864	3 802 792	3 825.90
阶段活动关系	54 740	2 015 388	3 581.70
物体功效关系	25 256	965 785	3 723.90
别称关系	5 758	16 354	184.00
因果关系	2 157	2 760	27.96
合计	194 775	7 275 958	3 635.50

表 16 句法模式有效性测试结果

Table 16 Validity test results of syntactic pattern

	句法模式	实体结点	实体关系
1000 篇测试数据中的种子概念数目	NA	1881	18278
人工句法模式	46	1580；准确率 84%	14256；准确率 78%
机器学习句法模式	63	1373；准确率 73%	12794；准确率 70%

结束语 专利文献分析的重点是技术特征的提取，这也是专利审查的最大难点，它是一个需要人工经验和耗时的工作。而仅仅通过 NLP 算法本身，不能有效地提升特征词的抽取准确率。通过本文描述的分领域构建句法模式进行特征词抽取，可以有效提升特征词抽取的准确率，继而有效提升专利文献的分析和审查效果。通过这种方法提取出来的特征词，结合知识图谱技术，可以有效地构建专利领域的知识库系统。这种可视化的知识库系统，一方面可以进一步推进专利审查和分析的准确率，另一方面可以成为专业和非专业人士学习专利领域知识的最佳途径。

参 考 文 献

[1] WIPO. World Intellectual Property Indicators 2021[R]. Geneva: WIPO, 2021.

[2] XU C L. Research method and application of technology development based on patent knowledge graph[D]. Guangzhou: South China University of Technology, 2017.

[3] XU J. Research on anti-liver cancer drug development trend in China based on knowledge graph and patent map[J]. Medical Information, 2018, 31(21): 19-23.

[4] SUN D. Research on patent measurement and knowledge graph in cloud computing field[J]. Sci-Tech Information Development & Economy, 2018, 3(6): 35-41.

[5] ZHANG Y, PAN H Q, LIN H G. Research on patent information of radix pseudostellariae based on scientific konwledge graph[J]. Journal of Anhui Agricultural Sciences, 2019, 47(6): 234-239.

[6] GAO S Y. Knowledge graph of Mongolian medicine patent in China: Citespace based metrological analysis[J]. Inner Mongolia Science technology & Economy, 2020(4): 96-101.

[7] ZHANG P L. Design and implementation of patent recommendation system based on knowledge graph[D]. Jinan: Shandong University, 2019.

[8] SERHAD S, LUO J X, KRISTIN L. Technology Knowledge Graph Based on Patent Data[J]. arXiv: 1906.00411, 2019.

[9] JI S X, PAN S R, ERIK C, et al. A Survey on Knowledge Graphs: Representation, Acquisition and Applications[J]. arXiv: 2002.00388, 2020.

[10] ZHANG N Y, DENG S M, SUN Z L, et al. Long-tail Relation Extraction via Knowledge Graph Embeddings and Graph Convolution Networks[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 3016-3025.

[11] NAYYERI M, CIL G M, VAHDATI S, et al. Link Prediction of Weighted Triples for Knowledge Graph Completion Within the Scholarly Domain[J]. IEEE Access, 2021, 8: 79521-79540.

[12] YAN C, SU Q, WANG J. MoGCN: Mixture of Gated Convolutional Neural Network for Named Entity Recognition of Chinese Historical Texts[J]. IEEE Access, 2020, 8: 181629-181639.

[13] YAN Z, PENG R, WANG Y, et al. CTEA: Context and Topic Enhanced Entity Alignment for Knowledge Graphs[J]. Neurocomputing, 2020, 410(3): 155-165.

[14] CHRISTINA L, THOMAS L, PATRICIA S, et al. Is buttercup a kind of cup? Hyponymy and semantic transparency in compound words[J]. Journal of Memory, 2020, 113: 104110.

[15] CHEN S D, OUYANG X Y. A review of named entity recognition technology[J]. Radio Communications Technology, 2020, 46(3): 251-260.

[16] YU T, CUI M, LI H Y, et al. Application research of ISO technical specification" Semantic Network of Chinese Medicine Language System"[J]. China Medical Herald, 2016, 13(4): 89-92.

[17] JI G L, HE S J, XU L H, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2015: 687-696.

DENG Liang, born in 1980, postgraduate. His main research interests include deep learning and knowledge graph.

CAO Cun-gen, born in 1964, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include large-scale knowledge process and so on.

（责任编辑：何杨）